Cultural bias and cultural alignment of large language models

Year: 2023 | Citations: 185 | Authors: Yan Tao, Olga Viberg, Ryan S. Baker, René F. Kizilcec

Abstract

Abstract Culture fundamentally shapes people’s reasoning, behavior, and communication. As people increasingly use generative artificial intelligence (AI) to expedite and automate personal and professional tasks, cultural values embedded in AI models may bias people’s authentic expression and contribute to the dominance of certain cultures. We conduct a disaggregated evaluation of cultural bias for five widely used large language models (OpenAI’s GPT-4o/4-turbo/4/3.5-turbo/3) by comparing the models’ responses to nationally representative survey data. All models exhibit cultural values resembling English-speaking and Protestant European countries. We test cultural prompting as a control strategy to increase cultural alignment for each country/territory. For later models (GPT-4, 4-turbo, 4o), this improves the cultural alignment of the models’ output for 71–81% of countries and territories. We suggest using cultural prompting and ongoing evaluation to reduce cultural bias in the output of generative AI.